Recitation 7: Machine Learning

Fall 2021

Find a partner!

What is Machine Learning?

“Machine Learning” has been a big buzzword in the past decade.

- Increase in computation power
- Increase in big data

Use of data to train a model that is able to do one of the following: prediction, image recognition, speech recognition, etc.

Scope of Machine Learning in 17-313

- High-level overview of what machine learning is
- Understand how to read data
- Have the technical know how to train a simple model
- Tinker with model parameters for better prediction scores

Main Focus: **Simple deployment of a ML model as a microservice.**

Goal of Recitation: **Prepare you for Homework 4.**

Starts with the Data!

Data is already collected :)

Understand your data and what it is about via data analysis.
Tech Stack (Python): pandas, sklearn, jupyter notebook, lime

Evaluate the important ***features*** to train your model on.

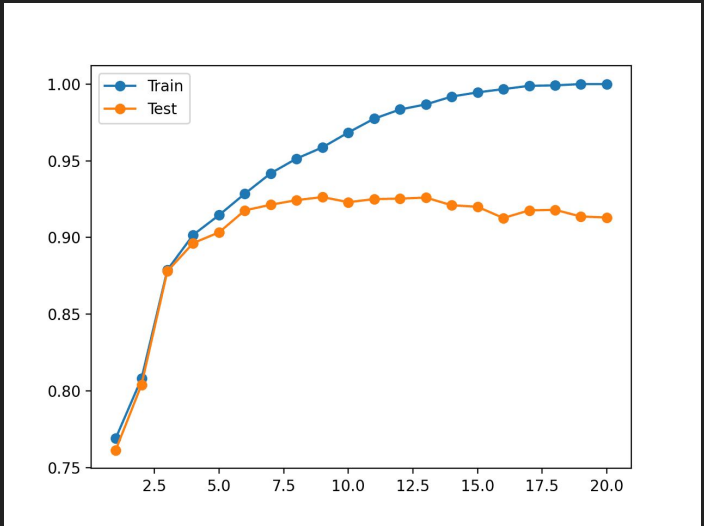
Preprocessing the Data:

- Remove outliers
- Add missing data points (Mean / Median)
- Feature Engineering

Train and Test Data

Split the data randomly into a training set vs. testing set.

- Helps determine ***overfitting***



pandas.read_csv

```
pandas.read_csv(filepath, sep=NoDefault.no_default,…)
```

Read in data from csv files to a pandas dataframe. Supports optionally iterating, breaking the files into chunks, and parsing data with some specified separators.

pandas.DataFrame.head

```
DataFrame.head(n=5)
```

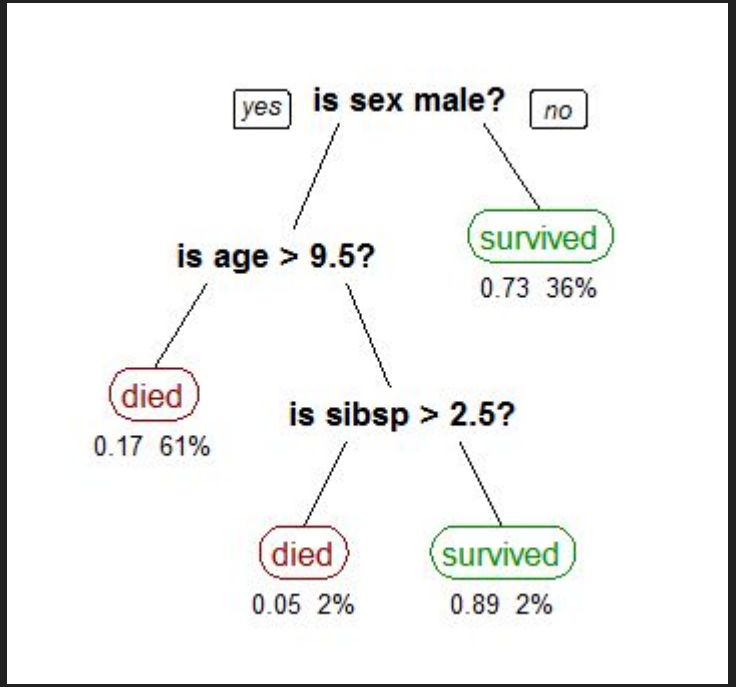
Return the first n rows.

This function returns the first n rows for the object based on position. It is useful for quickly testing if your object has the right type of data in it.

ML Models: Decision Tree

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions.

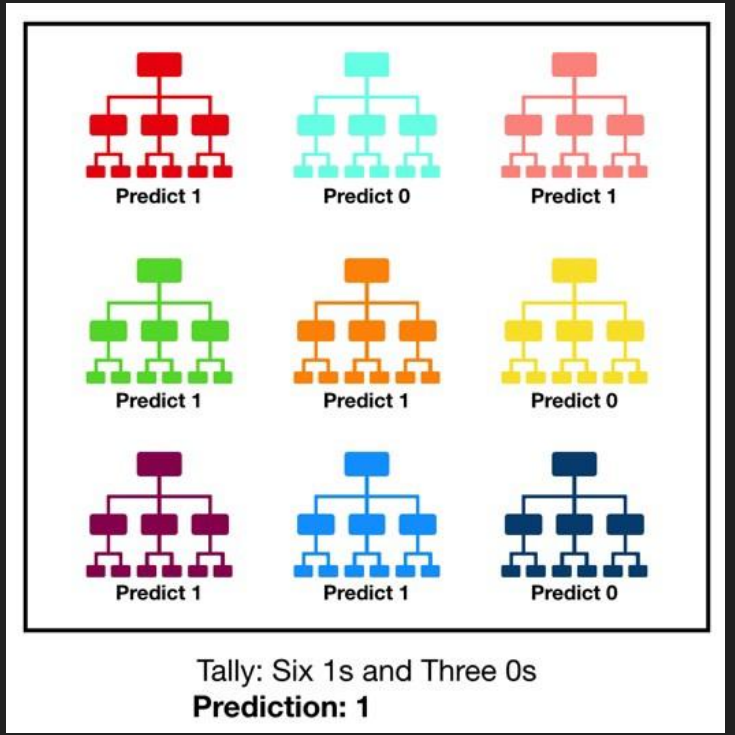
- nodes represent the asked question
- edges represent the answers to the question
- leaves represent actual output or class label



ML Models: Random Forest Classifier

Random Forest Algorithm consists of many decision trees.

- Predicts by taking the average or mean of the output from various trees.
- Increasing the number of trees increases the precision of the outcome



Activity 1: Examine Titanic Dataset

This dataset contains detailed information on the passengers aboard the Titanic. Our goal is to create a model able to predict whether a passenger will survive. However, before we start training our machine learning model, let us first explore the dataset. Use the pandas methods we went over earlier and explore what features are in the dataset.

Then choose one feature and explore its correlation with passenger survival rate. Hypothesize an explanation for why it has such effects.

* Repo: https://github.com/CMU-313/17313-recitation7-ML

Demo

Activity 2: Train Your Model

Using what we have learned earlier about Decision Trees and Random Forest Classifiers, work with your partner to train your own model to predict whether a passenger with given features will survive. Be sure calculate the accuracy of your model using the given test dataset.

We will have a mini-competition to find the best model!

* Feel free to use other methods to train your model.